国际 Data Curation 研究与实践发展综述

Review on the international development of research and practice of Data Curation

吴振新 [1] 陈瑶 [1,2] 李文燕 [1,2] 付鸿鹄 [1] 许丽媛 [1]

（[1] 中国科学院文献情报中心 北京 100190；[2] 中国科学院大学 北京 100190）

【关键词】 科研数据 研究数据 保管 保存 综述 挑战 发展机遇

Key words： Science Data Research Data Curation Preservation Review Challenges Development Opportunities

【摘 要】 通过调研国际主要机构的战略规划，归纳出 Data Curation 在管理、资源建设、技术基础设施方面存在的主要挑战。针对这些挑战，从战略规划、数据评估与遴选政策、关键技术、审计和认证四方面全面回顾了国际 Data Curation 研究、实践的发展情况。分析图书馆在大数据科研环境下可能参与科研数据保管的领域，为图书馆在 Data Curation 活动中谋求发展机会。

Abstract: The paper summarizes current challenges of Data Curation in management, resource development, technology infrastructure base on reviewing major research institutions' strategic plans；and fully reviews the developments of Data Curation research and practice about strategic plan, data appraisal and selection, key technologies, audit and certification against these challenges.The paper concludes potential domains that library can participate in Data Curation，tries to find more development opportunities for libraries in this area.

【中图分类号】 G203

1 前言

信息技术的发展引起了数据和信息容量的爆炸，也催生了新的科学研究模式－e-Science 的发展， Jim Gray 博士将这种新的数据探索型研究方式称为科学研究的“第四种范式”(The Fourth Paradigm)，这标志着科学研究从以计算为中心转变到以数据为中心，数据成为科研的灵魂。如今，数据已不再仅仅是收集和存储的对象，它已经转变成国家的基础战略资源，可以用这种资源来协同解决其他诸多领域的问题，如在马航 MH370 失联客机事件中，中国科学院遥感与数字地球研究所利用其保存的遥感卫星数据与马航失联客机疑似位置海域的卫星遥感数据对比，进而确定了几处油迹带区域，这对于客机失事位置的确定有重要作用。

Data Curation 常被译为数据保管、数据保存等，这里的 Data 主要是指科研数据。业内有很多有关 Data Curation 的定义，比较有代表性的是英国数字保管中心（Digital Curation Centre， 以下简称 DCC）的定义：“Data Curation 指的是在数字数据的生命周期内，对这些数据进行维护、保存以及实现价值增值的所有活动，这些活动能够提高现有数据的长期利用价值；主动管理这些数据有利于减少在重新研究这些数据时出现的各种威胁以及降低因数字技术的退化而带来的各种风险；同时，Data Curation 所进行的一些列活动还能使在

可信赖仓储库中保管的数据能够更广泛地共享给研究机构，以便支撑未来的研究活动。”[1]

Data Curation 是 e- Science 环境下科研数据共享和大规模科学计算的产物，是应对“大数据”时代科研数据管理和保存需求的一种必然的管理实践。中国科学院文献情报中心为了顺应新的科研服务需求、建立有效服务于数字科研的新模式、新功能、新机制，开展了系列战略调研活动，本文为“数据管理和数据基础设施建设新技术方法”部分成果。

2 Data Curation 面临的挑战分析

随着科研数据规模的扩大和数据种类的不断增加，传统的数据保存方法已无法满足需求，虽然越来越多的机构不同程度的涉足 Data Curation，但作为一个新兴的研究领域，Data Curation 依旧面临着许多问题及挑战。

美国国家数字管理联盟（National Digital Stewardship Alliance，以下简称 NDSA）在其 2015 年日程中将目前在数据管理领域遇到的问题及挑战归纳为以下几点：1）建设数字内容集合的关键问题，包括数字内容的全局性问题、大规模内容选择的方法、特殊格式数字内容的挑战；2）缺乏支持保存活动的资源导致对于成本、价值的研究需求增强；3）缺乏足够的数字管理人员；4）技术基础设施的发展方面，包括协调分布式服务生态系统亟待发展、制定文件格式行动方案、内容完整性的保障等[2]。

UK Data Archive 在其 2010—2015 年战略规划中将 Data Curation 面临的挑战归纳为：1）建立和颁发存储认证；2）规划方面需要保证多来源的资助、与用户的期望和技术要求同步及推进合作发展 ；3）建立更有效的管理结构和内部记录管理系统；4）开发有效的数据（集合）选择、采集、摄入和保存的工具，提高数据质量和数据包的有效性，发展自存档；5）开发新的数据访问模式、分发和可视化工具，重构数据注册和许可系统，整合相关数据服务[3]。

DCC 则将未来数十年 Data Curation 将遇到的挑战归纳为以下几方面：1）数据管理软件的发展；2）数据管理计划中审核承诺的一致性；3）有限期的数据保存策略的影响（管理评估）；4）明确应保存的数据资源；5）数据的知识产权；6）理解真正语义上的长期保存[4]。

可以看出，未来一段时间内的 Data Curation 所遇到的挑战和问题集中在以下几个方面：管理方面主要集中在战略规划、成本研究、人员与培训教育、知识产权、审计与认证等方面；资源建设方面主要集中在大规模的数据选择与特殊格式的资源保存；技术基础设施建设主要集中在数据组织、格式管理、数据的质量保障（完整性保障)、保存系统（工具）及体系架构的发展等方面。

3 Data Curation 研究与实践发展

近年来，许多机构、项目在 Data Curation 领域展开了大量深入的研究与实践。本文基于上述有关 Data Curation 面临的挑战，初步总结和分析各机构和项目为应对这些挑战所开展的相关研究和实践活动。由于作者研究领域所限，本文没有涉及教育培训和知识产权方面的研究。

3.1 战略及规划

战略及规划的制定是开展 Data Curation 首先要解决的重要问题，这些战略规划包括政策规划、可持续发展战略、合作战略等方面。目前，国际上对 Data Curation 在全局政策规划、合作战略的研究比较成熟，出现了一些具有实践参考价值的战略框架、解决方案以及工具，但可持续发展战略的研究还处于起始阶段，仅在成本研究上出现了少数研究成果，还不足以支持保存实践活动。

3.1.1 Data Curation 政策规划

在政策规划方面，DCC 提供了大量参考资料和行动指南，并提供了一个有关制定研究数据管理策略的方案[5]，它包括五个步骤：1）列出现有的管理框架；2）制定一张管理内容的表格；3）获得管理者的支持；4）咨询、起草及修改；5）批准与实施。

MaRDI-Gross 项目也给出了在“大科学”背景下制定数字管理规划（Digital Management Plan，以下简称 DMP）的解决方案[6]，它从 1）制定保存目标；2）数据发布计划；3）数据验证；4）软件及服务的保存；5）成本及成本模型；6）数据丢失模型化六个方面来制定 DMP 的实践流程框架。

目前，已有成型的 Data Curation 规划制定工具可供使用，包括 DCC 开发的 DMPonline、UC3 开发的 DMPtool、IDMP 开发的 CARDIO、SCAPE 开发的 Plato 以及 OpenDOAR。

3.1.2 合作战略规划

数据体量的指数增长和数据类型不断的复杂化，给 Data Curation 带来了越来越严峻的挑战，为解决 Data Curation 面临的问题和减轻保存风险，跨领域的合作行动计划的需求不断增加。

DCU（Digital Curation Unit）通过推动跨学科合作研究规划和行动计划来帮助解决 Data Curation 问题，它提出了一个包含六方面的行动计划[7]：1) 用生命周期的方法来管理保管的信息对象，其中应包含与指定社团的动态互动；2) 采用以事件为中心的方法，充分表示数据的“活动事件”；3) 广义上的 Data Curation 实践者应包括那些参与生成信息对象的公共传播及利用的相关人员；4) 确定一个基本的跨学科范围，使 Data Curation 能充分满足学科差异化需求；5) 使信息对象的相关解释性内容作为社区的数字记忆，并进行模拟存档；6) 提倡面向机构的方式来保管。

随着合作政策的发展，一系列有效的合作实践在数字保管的各个方面都产生了积极的影响，如促进开源软件开发的协作、人员和资源信息的共享、参与标准和实践的开发、协调数字保管责任、开发协作的遴选决策和数字集合政策等。在这方面表现突出的有国际互联网保存联盟（International Internet Preservation Consortium，IIPC），其成员合作开发了一系列开源工具，并支持可持续的共享维护模型。

同时，有关的合作组织机构不断增加，如全球 CLOCKSS 网络，它通过分散的、地理间不同的保存模式来确保组织内共同的数字资产得以完整地保存；Data—PASS 是一个自愿的机构组织同盟，目的是为了存档、编目、保存社会科学研究使用到的数据；MetaArchive 是由众多的记忆机构组织和创建的数字保存网络，同时也是一个安全且具有成本效益的仓储；DPN（The Digital Preservation Network）长期保存网络通过在不同的节点上保存数据集

的副本来防止由于技术、组织或自然灾害等原因而导致的灾难性损失。这些组织和他们所示范的多机构管理方法在使用和社会认可度方面均显著增加。

3.1.3 可持续发展规划

完成数字管理任务需要适当的资源来支持，但不可能有足够的资源来支持存储机构保存所有的数据，如何有效地对保管成本进行预算、管理及分配以及如何获得所需的资源已经成为可持续发展的重要问题。但由于 Data Curation 本身的复杂性及涉及多方利益，数字管理成本估计比较复杂和模糊，目前几乎没有模型能支持成本估算的比较数据或纵向数据。

4C（Collaboration to Clarify the Costs of Curation）是欧盟资助的主要致力解决保存费用问题的项目，他们分析了现有的 10 种成本模型及工具，并对每一种模型进行了分析及评价，通过分析已有的数字保存成本建模工作，他们提出了建立可持续性数字保存和获取的最佳实践建议。目前，4C 提供了一个包括尝试解决效益、风险、价值、质量和可持续性的成本模型工具和框架，并初步制定了一个经济可持续性参考模型、开发了一个保管成本交换平台工具—CCEx。

POWRR 项目则是利用有限资源进行数字对象长期保存研究的重要项目。它旨在帮助那些因缺少资源而难以开展数字保管的中小型机构。该项目正在评估能够在中小机构中实现数字长期保存的工具和服务，以期提供有效的解决方案。

这些项目的成果将有助于厘清成本以及辅助决策和战略规划的制定，反过来也可以促进数字保存的长期管理和发展可持续的基础设施建设。

3.2 数据评估与遴选政策

数字数据的特征使得对它的收集变得异常复杂并因此在保存方面也变得复杂。数据规模一直在扩大，数据的粒度和互联性也变得更加繁杂。传统的资源评估和遴选通常会基于机构自身的优先级、能力和指导政策，而数字数据则有其特殊性，使得相应的数据评估和遴选政策也更加复杂化。

NDSA 提出了一系列有关数据评估和选择的推荐做法，包括数据相关性、文档、资金、研究和应用的需求、可用性、风险和易用性等方面，这将有助于机构启动涉及整个信息生命周期的数字管理计划。

DCC 提出了一个选择及评估保管数据的方案[8]，即通过一个弱分析框架来辅助决定需要保管的数据，其中要考虑的因素包括：1）难以评估未来重用价值的数据； 2）学科形成前的数据； 3）数据及相关文档的质量； 4）不可替代的观测性数据（与实验数据相对）；5）重新生成实验数据的成本；6）估算保存具体数据集的成本。

NERC（Natural Environment Research Council）于 2012 年发布了数据权重清单（NERC Data Value Checklist），以便科研社区选择需要保存的数据。

研究实践表明，目前渗透到生活、文化及学术各方面的大量数字数据还无法被图书馆或档案馆获取，因此在遴选政策中应优先收集这样的原生数字材料，同时应积极获取特殊的原生数字材料（如网络档案、数字记录、文档及手稿档案的硬盘等），另外对数字材料的选择经常与机构的实力和使命相关。

3.3 Data Curation 的关键技术发展概述

3.3.1 元数据标准规范的制定和形成

元数据一直是 Data Curation 关注的重要领域。许多著名的机构和项目都推出了自身的元数据标准或推荐规范。

NDSA 的“数字保存级别”定义的四个级别包含了 Data Curation 流程中的不同元数据，分别是记录型、管理型、描述型、结构型、技术型元数据以及保存元数据。

DCC 发布的关于学科元数据标准的相关信息（元数据的概念、使用群体和使用方法）引起了研究数据管理（Research Data Management，RDM）社区的极大关注，随后专门创建了学科元数据网页[9]以帮助那些需要确定采用哪种元数据标准满足自己需求的用户。

韦恩州立大学提出了用于文物数字保存的语境元数据框架，这个框架由八个语境维度组成，并对需要捕捉的信息类型进行了识别，该框架可确保在一个元数据方案中记录充足的语境信息，从而为将来的搜索、检查、利用、管理和保存活动提供极大的便利。

Research Data @ Essex 以 IDMB 项目的一个元数据模型为出发点，建立了一个三层元数据模型。

2013 年 4 月，英国公布了一个用于该国存储库的元数据应用纲要和指导原则（RIOXX）。

美国声音记录元数据方案开发项目为其记录的音乐制定了一个用于收集和管理元数据的标准方法并开发了一个工具（Content Creator Data Tool，CCD）来帮助数据产生者及拥有者收集数据。

3.3.2 文件格式的识别、选择与转换

数字文件格式的稳定性和文件格式过时的风险是数字管理机构的重大挑战，特别是在大数据科研环境下，如何选择一种好的数据格式来保管数据是一项有挑战性、前瞻性的任务。面对正在积累大量的数字集，切实可行的、用于监测和挖掘机构所管理的异质原生数字文档的信息的策略和手段尤为重要。

欧洲聚变发展协议（European Fusion Development Agreement，EFDA）为了防止文件格式过时，在 Data Curation 实践中对如何选择文件格式提出了明确的解决方案[10]，即保管机构应该保存所有使用到的文件格式的核心信息并记录这些文件格式用到了哪些数据上，且这些核心信息应该经常更新；当选择一种格式用于 Data Curation 时，仅仅考虑到这种数据格式的当前表远远不够，还应该考虑到数据格式的长期性及未来的发展潜力。

美国国家档案馆和记录管理局出台的《公开发布的格式行动方案》通过鼓励数字内容产生部门去选择一组更精确的数字化格式来推动实践的发展，尤其像能在一定程度上实现集中控制的部门，如联邦、州、地方和区域政府。

NDIIPP 支持的“地理空间归档和保存合作计划（GeoMAPP）”项目的地理空间数据文件格式参考指南提供了一个关于一些常见的地理空间栅格数据与矢量数据集类型的快速参考，并且成为快速确定州政府常见的地理空间的文件格式类型的服务工具。

NDSA 最近发布了对 PDF/A 格式标准的研究报告，报告分析了曾经作为长期保存的黄金标准格式之一的 PDF/A 的特性以及对长期保存的影响。

美国国会图书馆发布了长期保存的推荐格式规范，FDA (Florida Digital Archive)也发布了自己的格式选择范围。Archivematica 在其软件平台上将格式策略和行动计划转化为由工具和软件直接实施和管理的行动，在实践上率先迈出了至关重要的一步。

相关可利用的工具包括：英国国家档案馆的文件格式管理工具系统 PRONOM、全球文件格式注册系统 GDFR(Global Digital Format Registry)。用于格式识别、校验、特征抽取的开源工具包括：JHOVE(LGPL)、DROID、用于文档格式受损分析的 Fuzzy Logic 以及相关的规范 PDF 验证工具和方法。

3.3.3 数据不变性和完整性的验证

Data Curation 中最重要的任务之一是保证数据的不变性和完整性，数据验证对确保数据可信发挥着重要作用。常用的验证数据不变性与完整性的方法是检查数据的不变性信息（Fixity Information），它能检测数据是否已遭破坏、监控硬件的退化、满足可信赖需求（如 ISO 16363/TRAC、NDSA 的数字保存级别）、支持文档起源和保管链、帮助诊断在 Data Curation 的管理周期中可能出现的系统或人为错误等等。

不变性检查通常分为两大类：1）统计性不变性检查，以统计文档数量和文件大小来进行不变性检查；2）内容不变性检查，多采用算法通过对文档内容进行比较和计算来进行不变性检查，以确定文档内容是否发生改变。

斯坦福大学的 LOCKSS 系统使用了 Opinion polls 机制，即利用保存同样内容的多个结点来进行定期的内容比较和监控。

Fedora Repository 则使用 MD5 来验证数字对象的不变性，Fedora 会为每个存档对象的数据流（Datastream）片段及其每个版本生成并保存 MD5，以方便进行数字对象的不变性校验。

DAITSS 系统利用 MD5 和 SHA1 算法定期计算全部文档副本的校验码。

UC3 的 Merritt 仓储库以微服务的方式提供多种类型的接口，并支持各种常用的摘要类型，可通过配置服务可以在任意时间实施不变性验证。

常用于产生与核查不变性信息的工具和算法有：Expected File Size、Expected File Count、CRC、MD5、SHA1、SHA25。目前专门为长期保存而开发的不变性、完整性工具有马里兰大学 ADAPT 项目开发的开源工具 ACE（Auditing Control Environment）和正在开发的用于验证数据集的本体工具 vplan。

3.3.4 数据唯一标识符与数据注册

如何对庞大的数据进行唯一标识是 Data Curation 机构面临的一个关键问题，保管人员选择采用通用的标识符体系来与传统资源保持一致，包括 ARK（持久标识符架构）、DOI（数字对象标识符）、Handle（句柄系统标识符）、URN（统一资源名称）、PURL（持久统一资源定位符）、URI（统一资源定位符)等。

同时也出现了专门的研究数据注册服务，ANDS 的 Cite My Data 服务能帮助研究机构为被引用的研究数据集自动分配 DOI。此外，为数据分配标识符服务的系统还有大英图书馆开发的 DataCite、UC3 开发的 EZID、WebCite 等。

3.3.5 保存技术策略

多年的保存研究和实践中逐渐形成了多种多样的、更符合实践需求的应用型的技术策略，作者曾进行了详细的介绍和评述[11]，本文仅对后续发展情况进行相应的补充。

比特保存通常被认为是最简单、最好理解的保存方法而被普遍所采用；格式转换和迁移也是目前被很多项目所采用的一项有效的技术策略；而仿真则是被认为未来最有效的保证数据可用性的重要措施，但由于其投资需求大、技术难度大、使用门槛高，目前只有少数项目在开展相关研究。

欧盟第七框架支持的 KEEP 项目提出了 “仿真作为服务”的方法，其发布的仿真框架（Emulation Framework）允许用户利用仿真来访问旧的计算及文件和程序，目前已经应用于 CD 数据以及 Web 信息的仿真服务。

SCAPE 项目则在基于格式迁移、格式风险、存储库性能的证据基础研究上开展了大量工作。

3.3.6 大规模数据保存的系统与基础架构

急剧增长的海量数据、数据对象（集合）更新的速度（频率）以及数据对象的多样性（异质性）给大规模的数据保存系统与基础架构带来了巨大挑战。

SCAPE 项目主要致力于解决密集型计算、保存平台可扩展性的问题，它分为大规模数字归档、科学数据集和网络归档三个子项目展开研究，主要处理科学数据和科学工作流。在应对大数据的挑战方面，SCAPE 已经初见成果，提供了基于实践的解决方案，构建了以数据为中心的分布式的 SCAPE 长期保存平台，可以为大型数据的执行过程提供基础设施。

UC3 面向大数据存储的 Merritt 系统通过采用“微服务（micro-services）”的开发模式,使得系统的规模和功能能够以微服务这种模块化模式扩展和更新,微服务小而独立的特点使它们更容易开发、部署、维护和升级,使得 Merritt 具备了大数据保存系统的理想特征，如服务高可用性、高可靠性、高效率、适应性和可持续性等。

斯坦福大学的 LOCKSS 系统采用的是典型的分布式存储方式，它为图书馆提供的是一个开放性源码的分布式存储系统，可以在本地收藏、管理电子资源。LOCKSS 利用多机构参与、多副本存储的机制，实现大量数字资源的可靠保存。

由 SDSC、加州大学圣地亚哥分校图书馆、美国国家大气研究中心(NCAR)和马里兰大学等合作的 Chronopolis 则提供了美国最大规模的协作式保存环境，利用网格技术在多站点和多平台间提供海量数据的监控、维护和存档管理。

Archive-It 是一个非营利项目-互联网档案馆（Internet Archive）的网络存档服务，它帮助机构获取、构建和保存数字内容集合。

Portico 是由世界上最大的数字存档社区所支持的数字存档，它能提供一个可持续性的业务模型来帮助图书馆、出版商和资助者协作保存电子期刊、电子书等电子学术内容。

DuraCloud 服务则以一种经济高效的代理方式利用众多的云存储提供商（包括商业及非盈利）为图书馆和研究机构解决了数字内容安全存储的基础设施问题。

3.3.7 小结

从上述可以看出，关键技术发展一直是 Data Curation 在推进过程中的重要研究和发展主题，经过多年努力，Data Curation 在关键技术的研究实践上取得了较为丰硕的成果。

在元数据的标准制定方面，很多项目基于已有的标准规范相继提出和定义了一些满足数据保管特殊需求的元数据框架和规范，这种集成和融汇的做法更有利于保证快速满足保存实践的需求，同时也能确保元数据标准的可用性；格式管理，作为保存中非常重要的一项工作，已经有多个机构推出了不同类型数据的适于保存的推荐格式集合，同时出现了很多开源的格式校验工具，并通过格式注册等机制来共同解决格式过时以及格式转换的问题，是相对发展较为成熟的领域；数据完整性检验作为保障数据长时间真实可用的有效手段，Data Curation 领域则是采用现有成熟的技术方法，通过制定针对实际需求的整体机制来予以解决；保存技术策略属于近几年来投入和研究较少的领域，只有少数项目针对仿真技术开展深入研究，其他研究甚少；而为了应对不断扩大的数据规模，很多机构探索和开发了不少适合于大规模数据保存、具备灵活可扩展特性的系统与基础架构，从各种角度和各种层面力图解决数字存储的基本问题。

3.4 审计与认证的发展

经过近年来的蓬勃发展，Data Curation 的审计与认证的研究与实践取得了一定的进展，许多可信赖的内容管理工作过程都得到了认可和标准化，同时也形成了一些国际标准。

RLG 在 2007 年发布的《可信赖仓储的审计及认证：指标与列表（Trustworthy Repositories Audit & Certification: Criteria and Checklist, TRAC ）》于 2009 年成为 ISO 国际标准（ISO 16363）。德国 nestor 制定的《可信赖数字仓储的指标体系》于 2011 年成为德国国家标准。荷兰 DANS 项目开展了数字认可证明授予服务，提供了 16 个指导方针供仓储库进行自评估。

欧盟则在上述三个标准规范的基础上提出了包括基本认证（依据 DSA 进行自评估）、扩展认证（依据 ISO 16363 或 DIN 31644 进行有组织的外部审计，提供公开的自评估）、正式认证（依据 ISO 16363 或 DIN 31644 进行全面认证）的三层认证框架。

DCC 以 TRAC 与 nestor 指标为基础，并在其中引入风险管理的概念，开发出一套"基于风险管理的数字仓储审计方法"（Digital Repository Audit Method Based On Risk Assessment，DRAMBORA）。

澳大利亚国家和州图书馆（National and State Libraries Australasia，NSLA) 为了评估成员馆的长期保存活动，基于美国卡内基梅隆大学的软件能力成熟度模型(capability maturity model ，CMM)提出了一个包括初始、可重复、定义、管理、优化等 5 层保存能力成熟度模型。

Tessella 公司为了协助开展长期保存的机构选择长期保存解决方案，提出了数字存档成熟度模型（Digital Preservation Maturity Model），用于识别不同类型的长期保存解决方案的成熟度。

NDSA 发布的"数字保存级别"是一套分层次的技术实践指南，旨在为保存数字内容提供清晰的技术基准说明，同时允许机构对他们保管的特殊资源进行保存级别评估。

尽管已有许多的研究、实践成果，但仍有许多工作要做，目前还没有保存社区广泛认可的认证过程。而针对集中式和分布式保存网络的可靠性研究刚刚起步，开发出一个全面、健硕的保存网络信任框架依旧是一个重大挑战。

4、结语

数据带来了科学研究范式的革命性变化，科研数据保管也为图书馆开展新的服务带来了机会与挑战。图书馆不仅可以主动参与到 e-Science 环境中，更可以凭借自身的优势为科研数据的保管提供重要支持。霍普金斯大学图书馆馆长 Winston Tabb 认为："e-Science 环境下，图书馆是分布式网络的一部分、数据能够成为馆藏资源、数据中心会成为新型图书馆书库、图书馆员是数据科学家并能提供数据服务。"[12]图书馆应顺应需求、抓住机会，打造有效服务于数字科研的新模式、新功能、新机制。

图书馆可以基于科研数据生命周期，研究探讨大数据科研环境下的科研数据保存管理的解决方案。主要研究包括：

- 科研数据保管规划研究

每个科研机构都需要根据实际需求制定自己的 Data Curation 政策，以此明确自身在科研 Data Curation 中的职责，并将政策作为一个执行框架来指导具体的研究 Data Curation 行动，包括数据遴选政策等。

- 合作模式与共享机制研究

Data Curation 行动应依据科研数据生命周期规律，与科研活动紧密结合，无缝嵌入科研流程，从而有效地支持并促进科研成果的产出、创新和共享。因此需要构建无缝嵌入科研流程的、与科研团队紧密合作的长期合作和共享机制。如何在尊重知识产权、符合政策法规的前提下进行有效的合作共享，将涉及政策、法规、技术等多方面问题，相关的政策激励、科研数据的版权和隐私保护是合作共享机制中必须考虑的重要问题。

- 服务内容及服务机制研究

研究在科研数据生命周期的各阶段所需要的保管服务内容，分析以怎样的方式无缝嵌入科研流程，以更加有效的方法提供多样化保管服务，使得科学数据能够发挥最大的科研价值、经济价值和社会价值，深入探索图书馆嵌入科研流程的、动态的科学数据服务机制与模式。

- 基础设施和关键技术研究

全面分析国际科研 Data Curation 基础设施（Research Data Curation Infrastructure，RDCI）方面的重要规划、进展、方案、技术框架和相关技术方法。特别研究文献信息机构介入 RDCI 建设的策略和业务模式，为融入科研生命周期的科研数据支撑和服务环境建设提供有益借鉴。深入研究研究 Data Curation 的关键技术方法，分析相关标准规范、技术策略和工具系统，构建大数据科研环境下的科研 Data Curation 技术框架。

- 素养教育研究

系统分析科研 Data Curation 和服务领域中各种角色（创造者、专家、管理者、数据馆员）的作用和职责，构建各种角色参与科研数据管理和服务所需的知识能力结构，为相关人员的培训和继续教育提供理论依据和教学材料框架。

- 可持续发展研究

详细研究覆盖研究 Data Curation 生命周期的成本与效益的模型，分析不同利益相关方需求和所负担的费用以及可获得的收益，为研究 Data Curation 活动确立和维持主要的投资提供具体的成本-效益分析；在此基础上进行可持续发展的经济模式研究，形成具有自我生存能力的研究 Data Curation 生态环境。

参考文献：

[1] DCC.What is digitalcuration?[EB/OL].[2014-12-2]. http://www.dcc.ac.uk/digital-curation/what-digital-curation.

[2] NDSA.2015 National Agenda for Digital Stewardship[EB/OL].[2014-12-2]. http://www.digitalpreservation.gov:8081/ndsa/documents/2015NationalAgenda.pdf.

[3] UK Data Archive.UK Data Archive Strategic Plan, 2010-2015[EB/OL].[2014-12-2]. http://www.data-archive.ac.uk/media/196518/ukda-strategicplan20102015full.pdf.

[4] Research Data Management: Practical Strategies for Information Professionals[M]. Purdue University Press, 2014:399-406.

[5] DCC.Five Steps to Developing a Research Data Management Policy[EB/OL].[2014-12-2]. http://www.dcc.ac.uk/sites/default/files/documents/publications/DCC-FiveStepsToDevelopingAnRDMpolicy.pdf.

[6] DMP Planning for Big Science Projects.[R/OL].[2014-12-2].http://arxiv.org/pdf/1208.3754v1.pdf.

[7] DCU.Key challenges and strategies[EB/OL].[2014-12-2]. http://www.dcu.gr/index.php?p=dcu&lang=en&section=11.

[8] Whyte A, Wilson A. How to Appraise & Select Research Data for Curation[M]. Digital Curation Centre, 2010.

[9] DCC. Disciplinary Metadata[EB/OL].[2014-12-2].http://www.dcc.ac.uk/resources/metadata-standards.

[10] Layne R, Capel A, Cook N, et al. Long term preservation of scientific data: Lessons from jet and other domains[J]. Fusion Engineering and Design, 2012, 87(12): 2209-2212.

[11] 吴振新, 张智雄, 郭家义. 数字信息资源长期保存技术策略分析[J]. 现代图书情报技术, 2006 (4): 8-13.

[12] Reilly S, Schallier W, Schrimpf S, et al. Report on integration of data and publications[J]. 2011.

［作者简介］

吴振新　女，1968，中国科学院文献情报中心研究员，硕士研究生导师。

陈　瑶　男，1991，中国科学院文献情报中心，中国科学院大学硕士研究生。

李文燕　女，1989，中国科学院文献情报中心，中国科学院大学硕士研究生。

付鸿鹄　女，1976，中国科学院文献情报中心馆员。

许丽媛　女，1986，中国科学院文献情报中心馆员。

地址：北京市中关村北四环西路 33 号中国科学院文献情报中心信息系统部

邮编：100190

电话：15600602409

电子邮箱：chenyao@mail.las.ac.cn